Speech Recognition System for Windows Commands

Sumit Patel
Lecturer in Computer Dept,
SNJB's KBJ COE, Chandwad

Amit Bramhecha
Lecturer in I.T Dept,
SNJB's KBJ COE, Chandwad

Santosh Mahale
Lecturer in I.T Dept,
SNJB's KBJ COE, Chandwad,

Anant Maind
Assistant Prof in I.T.Dept
SNJB's KBJ COE, Chandwad

Mahesh Sanghavi
H.O.D of Computer Dept,
SNJB's KBJ COE, Chandwad,

ABSTRACT

To develop a system to recognize system commands through voice and convert it into equivalent text, the system accepts voice commands from user and displays its equivalent text. The system accepts voice commands, performs processing on it to recognize the actual command before displaying the corresponding output. For this particular system processing being done are noise removal, feature extraction and pattern matching. Various features are available. These are totally application dependent i.e. for a particular application particular feature is being extracted. Hence performing this various processing, text format of equivalent voice command is being displayed. To accept the voice commands User Must use a good quality microphone. The voice commands is being recorded and saved as a .wav file. Wav file is being used because it stores the data in the digital form. Initially the features of each command would be saved in a file. Once the 'init' is recognized the system will then wait for the users commands. On getting a command the system will save the input as a .wav file. The features of this command are then matched against the predefined command features. If the match is found the command is a valid one. It then displays its text form. If the command is not valid it simply discards it.

Keywords

Recognize, Feature Extraction, Pattern Matching, Noise Removal.

1. INTRODUCTION

Speech is one of the oldest and most natural means of information exchange between human beings. Humans speak and listen to each other in human-human interface. For centuries people have tried to develop machines that can understand and produce speech as humans do so naturally [2]. Attempts have been made to develop vocally interactive computers to realise voice/speech recognition. Voice/speech recognition is a field of computer science that deals with designing computer systems that recognize spoken words. It is a technology that allows a computer to identify the words that a person speaks into a microphone or telephone. Speech recognition can be defined as the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words [4].

Automatic speech recognition (ASR) is one of the fastest developing fields in the framework of speech science and engineering. As the new generation of computing technology, it comes as the next major innovation in man-machine interaction, after functionality of text-to-speech (TTS), supporting interactive voice response (IVR) systems. Nowadays, the statistical techniques prevail over ASR applications. Common speech recognition systems these days can recognize thousands of words. The last decade has witnessed dramatic improvement in speech recognition technology. In some cases, the transition from laboratory demonstration to commercial deployment has already begun [4].Aspects of our daily life, for example, telephone applications, applications for the physically handicapped and illiterates and many others in the area of computer science. Speech recognition is considered as an input as well as an output during the Human Computer Interaction (HCI) design. HCI involves the design implementation and evaluation of interactive systems in the context of the users' task and work [5].

2. Related Work

Human computer interactions as defined in the background is concerned about ways Users (humans) interact with the computers. Some users can interact with the computer using the traditional methods of a keyboard and mouse as the main input devices and the monitor as the main output device. Due to one reason or another some users cannot be able to interact with machines using a mouse and keyboard [6] device, hence the need for special devices. Speech recognition systems help users who in one way or the other cannot be able to use the traditional Input and Output (I/O) devices. For about four decades human beings have been dreaming of an "intelligent machine" which can master the natural speech [7]. In its simplest form, this machine should consist of two subsystems, namely automatic speech recognition (ASR) and speech understanding (SU). The goal of ASR is to transcribe natural speech while SU is to understand the meaning of the transcription.

2.1 Talkman play station portable

Talkman is the voice-activated multilingual translation software developed by Sony Computer Entertainment for the Sony PlayStation Portable video game console. Its first release offers translations, mostly slang and helpful travel phrases, between all four languages, i.e. Japanese, Chinese Mandarin, Korean, and English. Talkman also includes game mode to help players learning and practicing languages. The second release of Talkman is Talkman Euro which provides translations in six languages, that is, English, Italian, Spanish, German, French, and Japanese (or Traditional Chinese). According to the reviews of Talkman by Metacritic [23], it appears to be that Talkman receives an average or mixed score. The pros of Talkman involve its novelty and the ability to assist the users as a translator/tutor with interactive talking phrasebook and games functionality. However, the big cons of Talkman are its vocal translation technology and loading time.

The former is concerned with the problem of delivering the right phrases in response to the users' input, while the latter is about spending long time to switch between screens. In addition, Talkman does not include dictionary function either.To sum up briefly, it is noticeable that the commercial dictionary products presented above are now capable of speaking a headword and its definitions with Text-to-Speech technologies developed for their own company. Conversely, the products cannot be operated to look up word definitions by speaking a word, i.e. voice input. Only the Ectaco X8 series electronic dictionaries can accept voice input to search for word definitions but by spelling a letter not speaking a word, whereas Talkman translator accepts voice input to translate phrases into other foreign languages not for meanings lookup. Therefore, it is quite clear that the issue of using voice input by saying a word to look up word definitions opens an area for further development.

2.2 Human Interface Technology Laboratory

Speech recognition is the technology by which sounds, words or phrases spoken by humans are converted into electrical signals and these signals are transformed into coding patterns to which meaning has been assigned". While the concept could more generally be called "sound recognition" [14], focus here on the human speech because human most often and most naturally use his/her speech's to communicate these ideas to others in his/her immediate surroundings.

2.3 Exploring New Speech Recognition in Microsoft

Microsoft has been researching and developing speech technologies for over a decade. In 1993, the company hired Xuedong (XD) Huang, Fil Alleva, and Mei-Yuh Hwang—three of the four people responsible for the Carnegie Mellon University Sphinx-II speech recognition system, which achieved fame in the speech world in 1992 due to its unprecedented accuracy. Right from the start, with the formation of the Speech API (SAPI) 1.0 team in 1994, Microsoft was driven to create a speech technology that was both accurate and accessible to developers through a powerful API. The team has continued to grow and over the years has released a series of increasingly powerful speech platforms [24].

In recent years, Microsoft has placed an increasing emphasis on bringing speech technologies into mainstream usage. This focus has led to products such as Speech Server, which is used to implement speech-enabled telephony Systems, and Voice Command, which allows users to control Windows Mobile devices using speech commands. So it should come as no surprise that the speech team at Microsoft has been far from idle in the development of Windows Vista™. The strategy of coupling powerful speech technology with a powerful API has continued right through to Windows Vista. Windows Vista includes a built-in speech recognition user interface designed specifically for users who need to control Windows and enter text without using a keyboard or mouse. There is also a state-of the-art general purpose speech recognition engine. Not only is this an extremely accurate engine, but it's also available in a variety of languages. Windows Vista also includes the first of the new generation of speech synthesizers to come out of Microsoft, completely rewritten to take advantage of the latest techniques.

2.4 Speech Recognition Technologies

During an overview of speech recognition technology, software, development and applications the various parameters are considered. It begins with a description of how such systems work, and the level of accuracy that can be expected [16].

It describes:

- All users speak differently.
- How conventional speech recognition systems work?
- Reducing extraneous factors.
- Accuracy.
- Contemporary speech recognition systems.
- Dragon Naturally Speaking Preferred 6.0
- IBM ViaVoice 10.0
- Keystone SpeechMaster 5
- Microsoft Office XP (inbuilt)

2.4.1 Dragon Naturally Speaking Preferred 6.0

Adequate training for this package was found by reviewers to take around 20 minutes, or 10 minutes with follow-up practice. With NS, the user can "speak" into any open window, such as a Word 2002 document. The software also allows the user to "surf the web hands-free" - in a web page, if the user says the first few words of a link then the browser will go to the linked page. However, several reviewers indicated that the software requires a computer with a considerable amount of processing power to produce excellent results. As of January 2003, the package was available for roughly £124, with license deals available for multiple user use. A search of various websites indicated that versions of this package were the most frequently available speech recognition software in UK universities.

2.4.2 IBM ViaVoice 10.0

Training for this system takes between 20 and 40 minutes (depending on reviewer).Version 10 is very new, and according to some reviewers gives accuracy scores of over 96%, making it currently possibly the best package in terms of accuracy. Speech input, still resulting in a high level of accuracy, is possible up to a rate of 160 words per minute. However, the software requires a considerable amount of processing power to work very accurately, and over 0.5 of a gigabyte of hard drive space. As of January 2003, the package was available for around £80.

2.4.3 Keystone SpeechMaster 5

This package is a combination of Dragon NaturallySpeaking and Keystone Screen Speaker. The added functionality provided by the latter provides various spelling aids and word recognition support, making it suitable for people with Dyslexia. However, as of October 2002 the package costs £346, presenting a significant barrier to its use in school situations.

2.4.4 Microsoft Office XP (inbuilt)

Speech recognition is included within the latest version of the popular Office range of software. This allows the user to both enter commands, such as "open file", and to dictate text straight into an application. Added functionality enables the translation of text between a small numbers of languages. A small portion (four paragraphs) of this report was entered using this system. This proved frustrating at first, as the author became very conscious (and distracted) by the process of entering content, as opposed to concentrating on the content itself However, accuracy improved significantly from the first paragraph to the fourth.

2.5 Speech recognition research in the UK education sector

There are a number of research groups in the UK who are active in speech recognition, speech-to-text and/or text-to-speech research (though far less in the development of commercial products). There is unfortunately a lack of formal

analytical research into how effective speech recognition systems have been in UK.
Such research would have the benefits of:
• Proving or showing how effective such systems were
• Enabling fund-holders in similar educational institutions to determine whether to purchase such systems [15].

2.6 Voice Recognition Technology

Voice recognition technology or commonly referred to as speech recognition technology, is a constantly evolving type of technology. It has been experiencing tremendous growth in the commercial market. What it does is to translate human speech into electrical signals and then converts these signals into coding patterns with assigned meanings. This technology uses voice terminals that can be used for applications where operator's hands and eyes are occupied. Thus it enables data capture real time. Users typically wear a microphone/speaker headset like the one shown above, attached to a unit that recognizes spoken words. These words are then converted into analog electrical signals. The analog signals are converted to digital patterns, which are decoded by template-matching or feature analysis. The output data is entered into a program and may start a variety of computer-based equipment, for example scales, programmable logic controllers, or printers. In order to help this process comprising Hidden Markov Modeling is applied. This approach uses language models to determine how many different words are more apt to follow a particular word. The advantage is that groups of words that sound similar, for example "to", "two", and "too" are reduced and actual words are recognized. According to Ruggles, the error rates using this language modeling are from 1 to 15% [18].

2.8 ASR Engines from Academia

• Sphinx (CMU)
• HTK (Cambridge University)
• SUMMIT (MIT)
• SONIC (University of Colorado)
• Julius (CSRC, Japan)
• CSLU (OGI school of Science and Engineering)

2.9 Features of the proposed system

• Start Listen / Stop Listen.
• Accuracy.
• Mic Training Wizard.
• User Training Wizard.
• Change User Profile.

2.10 Summary of some commercial speech recognition systems

Speech Recognition Systems	Embedded Via Voice [25]	Open Speech Recognizer [27]	Windows Speech Recognition in Windows Vista [26]	Sphinx 4 [28]
Vocabulary and language				
Vocabulary size	Middle-size (500 words) to large (> 200,000 words) depending on selected software packages	Large	Normal to large	Small size (100 words) up to large size (64,000 words)
Extensibility	Dynamic	Dynamic	Fixed to changeable	Changeable
Usage conditions				
Environment:	Clean to Hostile	Normal to hostile	Clean to normal	No Information
Channel quality:	High-quality to low-quality	Normal to low-quality	High-quality to normal	No Information
Communication style				
Speaker:	Independent	Independent	Adaptive	Dependent and Independent supported
Speaking style:	Discrete and Continuous	Continuous	Continuous	Discrete and continuous
Overlap:	No Information	Barge-in	Barge-in	No Information

Table 2.1. The summary features of commercial and research speech recognition systems

3. Proposed System

The converted audio is next broken into phonemes by a phoneme recognition module. This module searches a sound-to-phoneme database for the phoneme that most closely matches the sound it heard. Each database entry contains a template that describes what a particular phoneme sounds like. As with text-to-speech, the table typically has several thousand entries. While the phoneme table could in practice they are different because the SR and TTS engines usually come from different vendors. Because comparing the audio data against several thousand phonemes takes a long time, the speech recognition engine contains a phoneme prediction module that reduces the number of candidates by predicting which phonemes are likely to occur in a particular context. For example, some phonemes rarely occur at the beginning of a word, such as the "ft" sound at the end of the word "raft." Other phonemes never occur in pairs. In English, an "f" sound never occurs before an "s" sound. But even with these optimizations, speech recognition still takes too long.

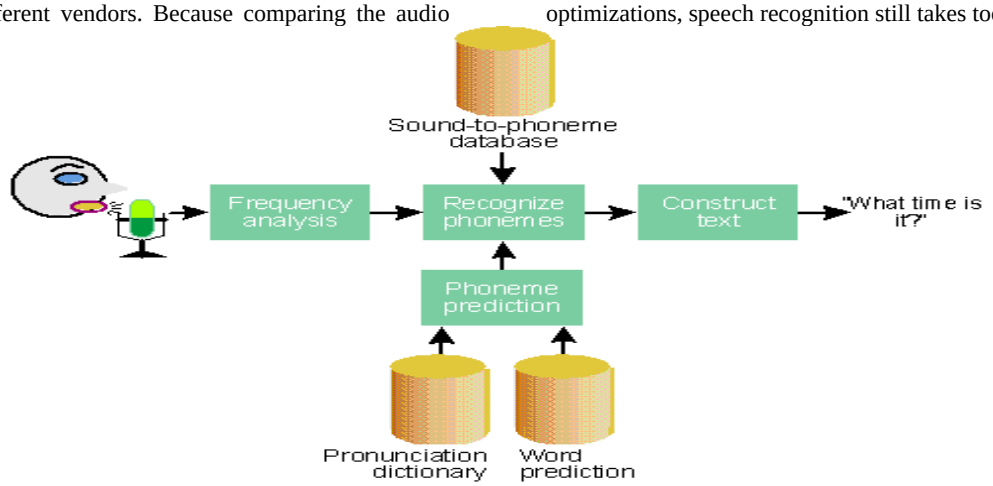

Fig 1. Conversion Speech to Text

A word prediction database is used to further reduce the phoneme candidate list by eliminating phonemes that don't produce valid words. After hearing, "y eh," the recognizer will listen for "s" and "n" since "yes" and "yen" are valid words. It will also listen for "m" in case user say "Yemen." It will not listen for "k" since "yek" is not a valid word. (Except in baby-talk, which is not currently supported?) The candidate list can be reduced even further if the application stipulates that it only expects certain words. If the app only wants to know if the user said "yes" or "no," the phoneme recognizer needn't listen for "n" following "y eh," even though "yen" is a word. This final stage reduces computation immensely and makes speech recognition feasible on a 33MHz 486 or equivalent PC. Once the phonemes are recognized, they are parsed into words, converted to text strings, and passed to the application.

Systems also need a sound card, microphone, and speakers. Most speech engines will work with any sound card. Some systems offload processing onto a DSP (digital signal processor) chip that comes on some high-end sound cards, which cuts the CPU speed requirement in half. Better microphones and speakers will also improve things.

As speech has become more feasible on average PCs, vendors have been busy developing and promoting their speech engines. Many multimedia PCs and sound cards come bundled with speech software. Others vendors sell their engines as standalone products. Some apps even come bundled with speech engines.

4. CONCLUSION

Speech recognition is the process of taking the spoken word as an input to a computer program. This process is important to virtual reality because it provides a fairly natural and intuitive way of controlling the simulation while allowing the user's hands to remain free. This partial project report will delve into the uses of Speech recognition in the field of virtual reality, examine how voice recognition is accomplished, and list the academic disciplines that are central to the understanding and advancement of Speech Recognition technology.

5. REFERENCES

[1] Markku Turunen and Jaakko Hakulinen, Design and Development of Speech Interfaces Course Material http://www.cs.uta.fi/hci/spi/ddsi/

[2] Pinker, S., (1994), the Language Instinct, Harper Collins, New York City, New York, USA.

[3] Deshmukh, N., Ganapathiraju, A, Picone J., (1999), Hierarchical Search for Large Vocabulary Conversational Speech Recognition. IEEE Signal Processing Magazine, 1(5):84-107.

[4] Zue, V., Cole, R., Ward, W. (1996). Speech Recognition. Survey of the State Of the Art in Human Language Technology. Kauii, Hawaii, USA.

[5] Dix, A.J., Finlay. Abowd, G., Beale, R. (1998). Human-Computer Interaction, 2nd edition, Prentice Hall, Englewood Cliffs, NJ, USA.

[6] Rudnicky, A.I., Lee, K.F., and Hauptmann, A.G. (1992) Survey of current Speech Technology. Communications of the ACM, 37(3):52-57.

[7] Picheny, M., (2002). Large vocabulary speech Recognition, 3 5(4):42-50.

[8] Rabiner, L., R., and Wilpon, J. G., (1979). Considerations In applying clustering Techniques to speaker-independent word recognition. Journal of Acoustic Society of America. 66(3):663-673.

[9] Kumar, M.Rajput, N.Verma, A .(2006) IBM Journal of Research and Development, 0018-8646,10.1147/rd.485.0703,Sponsored by: IBM

[10] De Mori, Renato, Lam, Lily, Gilloux, Michel. (1987) Pattern Issue, 0162- 8828, 10.1109/TPAMI.1987.4767902, IEEE Computer Society

[11] Bahl, Lalit R, Jelinek, Frederick, Mercer, Robert L, (2000), IBM T. J. Watson Research Center, Yorktown Heights, NY 10598. PAMI-5 Issue: 2 , IEEE Computer Society

[12] Liu, Y. Jones, H. Vaidya, S. Perrone, (2009). http://research.microsoft.com/pubs/80528/SPM-MINDS-I.pdf

[13] M.Tydlitat, B. Nanda, A. K. (2010), IBM Journal of Research and Development, Issue: 5, 0018-8646, 1147/rd.515.0583.

[14] Mengjie, Z., (2001) Overview of speech Recognition and related machine Learning techniques, Technical report. Retrieved December 10, 2004 from http://www.mcs.vuw.ac.nz/comp/Publications/archive/CS-TR-01/CS-TR- 01-15. Pdf

[15] "Research Developments and Directions in Speech Recognition and Understanding, Part 1" , (2009).http://research.microsoft.com/pubs/80528/SPM-MINDS-I.pdf

[16] Speech Recognition Technologies, (John Kirriemuir, 2003). http://www.ceangal.com/

[17] Speech Recognition – Wikipedia http://en.wikipedia.org/wiki/Speech_recognition

[18] Voice Recognition Technology http://cobweb.ecn.purdue.edu/~tanchoco/MHE/ADC-is/Voice/main.shtml

[19]http://www.opendl.net/solutions/recognition.aspx

[20]Casey Brains http://www.scribd.com/doc/6901516/ugSpeechSpeech

[21] Wolfgang Wahlster, Verbmobil: Foundations of Speech-To-SpeechTranslation http://books.google.com/books?hl=en&lr=&id=RiT0aAzeudkC&oi=fnd&pg=PR5&dq=Verbmobil:+Foundations+of+Speech-ToSpeech+Translation&ots=jBhMwQ0HnT&sig=zx2EWMK4n-lYhG9k5gKU2zGieE#PPP1,M1

[22] Roni Rosenfeld, Alexander Rudnicky, Stefanie Tomko, Thomas Harris, Universal Speech Interface project

http://www.cs.cmu.edu/~usi/

[23] Wikipedia the Free Encyclopedia – Talkman http://en.wikipedia.org/wiki/Talkman

[24] Talking Windows http://msdn.microsoft.com/da-dk/magazine/cc163663(en-us ,printer).aspx

[25] IBM Research, IBM Text-to-Speech Research http://www.research.ibm.com/tts/

[26] Microsoft Corporation, Windows Speech Recognition

http://www.microsoft.com/windows/products/windowsvista/features/details/speechrecognition.mspx

[27] Nuance Communications, Inc., Nuance – Open Speech Recognizer http://www.nuance.com/recognizer/openspeechrecognizer/

[28] Carnegie Mellon University, Sphinx-4 A Speech Recognizer Written http://cmusphinx.sourceforge.net/sphinx4/